Documenting Georeferenced Social Science Survey Data: Limits of Metadata Standards and Possible Solutions

Jünger, Stefan; Borschewski, Kerrin; Zenk-Möltgen, Wolfgang

Postprint / Postprint
Zeitschriftenartikel / journal article

This is the postprint of the article originally published in The Journal of Map & Geography Libraries: Advances in Geospatial Information, Collections & Archives, Vol. 15, 2019, No. 1, pp. 68-95 available online at:
https://doi.org/10.1080/15420353.2019.1659903

Documenting Georeferenced Social Science Survey Data: Limits of Metadata Standards and Possible Solutions

STEFAN JÜNGER, KERRIN BORSCHEWSKI AND WOLFGANG ZENK-MÖLTGEN
Data Archive for the Social Sciences, GESIS, Cologne, Germany

In this article, we present documentation of the georeferenced social science survey data that are spatially linked to geospatial data attributes. We introduce the challenges of documentation, as different metadata standards are used for both data sources: social science survey data and geospatial data. In particular, we analyze the extent to which the social sciences metadata standard DDI Lifecycle is capable of incorporating the geosciences metadata standard ISO 19115. We find that the most challenging attributes to describe are those concerning the geographic structure of the geospatial data, especially if they stem from different sources. To navigate these issues, we developed and evaluated four workaround approaches which we demonstrate in a case study on the georeferenced German General Social Survey. Because not all of the approaches apply equally to every research project and institution, we provide a scheme to assist in making informed and weighted decisions.

KEYWORDS social science survey data, georeferenced survey data, geospatial data, metadata standards, DDI lifecycle, ISO 19115

INTRODUCTION

In social science survey research, analyzing the context of social behavior is heavily supported by using areal information about respondents'

Address correspondence to Stefan Jünger, GESIS - Leibniz Institute for the Social Sciences, Data Archive for the Social Sciences, Unter Sachsenhausen 6-8, Cologne 50667, Germany. E-mail: stefan.juenger@gesis.org

neighborhoods, especially on a small scale (Nonnenmacher and Friedrichs 2011; Sluiter, Tolsma, and Scheepers 2015). By using small-scale georeferenced survey data, researchers can answer questions about individual social behavior or attitudes (Förster 2018; Termorshuizen, Braam, and van Ameijden 2015) while also taking into account the geospatial patterns of social processes (Klinger, Müller, and Schaeffer 2017; Legewie and Schaeffer 2016; Tolsma and van der Meer 2017). Consequently, in recent years, there has been an increasing demand to analyze (Hillmert, Hartung, and Weßling 2017; Meyer and Enzler 2013), access, and combine georeferenced survey data with other geospatial data (Bluemke et al. 2017; Jünger 2019; Schweers et al. 2016).

As data librarians, part of our work in this undertaking is to make sure that data are understandable and reusable, and we can accomplish this by getting involved in the early stages of a research project (Kong 2015). Usually, we use well-established metadata standards to achieve the goal of documentation in both the social sciences (Gómez, Méndez, and Hernández-Pérez 2016; Van den Eynden and Corti 2017; Jensen, Katsanidou, and Zenk-Möltgen 2011) and the geosciences (Porcal-Gonzalo 2015). Attributes of social science survey data depict information about respondents who are located in specific locations at the time of the interview. A well-established metadata standard for this type of data is the Data Documentation Initiative standard (DDI) (Vardigan 2013). Geospatial data, on the other hand, contain attributes that can constitute information on complex structures such as polygon or raster geometries. A well-established metadata standard in this field is, for example, the ISO 19115 standard (Ahonen-Rainio 2006) (https://www.iso.org/standard/53798.html). Thus, metadata standards already exist that are successfully used in the social sciences and the geosciences.

However, documenting research data becomes challenging if we aim to document data originating at the intersection of different scientific disciplines (Edwards et al. 2011), such as georeferenced survey data that are linked to geospatial data information. This linking implies a need to document data from different sources and of different types, and hence of different contents and different structures. Although we can in principle rely on the metadata standards mentioned, they were not designed to document linked datasets in all use cases.

In this article, we describe the specific use case of georeferenced survey data that are linked to different sources of geospatial data information, and we describe how to capture their metadata. In the following two sections, we provide some background on georeferenced survey data and spatial linking as well as background on the corresponding metadata standards, such as Dublin Core, DataCite, DDI, and ISO 19115, for both social science survey data and data from the geosciences. In the main section of this article, we discuss the challenges of integrating these standards

into one applicable scheme for our use case. We propose several solutions to navigate these challenges, demonstrate their necessity in a case study of the georeferenced German General Social Survey 2014, and discuss how specific considerations can provide criteria for choosing the appropriate solution for different projects. The last section concludes with a summary of this article but also expands the view on other data linking projects that would profit from metadata standards' ability to document linked datasets.

GEOREFERENCED SOCIAL SCIENCE SURVEY DATA AND SPATIAL LINKING

The Use of Georeferenced Survey Data in the Social Sciences

As mentioned in the beginning, in recent years georeferenced survey data has gained much attention with their promise to provide "new vehicles for innovation, synthesis and integration across the social and behavioral sciences" (Stimson 2014, 13). By drawing on methods of the geospatial sciences, social scientists can include characteristics of people's neighborhoods on a rather small scale, such as census grids, in their analysis. Moreover, these characteristics stem from a diverse set of other scientific disciplines such as ecology, engineering, or landscape planning. In addition to the potential to answer new and innovative research questions, georeferenced survey research is an emerging field of interdisciplinary discoveries (for a general overview please refer to Bluemke et al. 2017 or even more comprehensive to Jünger 2019).

This interdisciplinarity is also one of the strengths of georeferencing methods in social science survey research. New data that researchers add to existing survey data also offer new perspectives on existing research findings, and they add upon results that were limited through missing comparability. For example, a rather old question in the social sciences is how immigrant rates in the neighborhood affect prejudices towards foreigners among the native population (Allport 1954; Blumer 1958; Blalock 1967). By comparing immigrant rates in people's communities with bordering ones, authors in Sweden and Switzerland found that people in ethnically homogenous neighborhoods that are bordered by ethnically diverse neighborhoods feel more threatened by immigrants than people who live in ethnically diverse neighborhoods (Rydgren and Ruth 2013; Martig and Bernauer 2016). In Germany, however, there is no evidence for such an effect because of other residential segregation structures (Klinger, Müller, and Schaeffer 2017). Without the use of georeferenced survey data, revealing such international and contradictory comparisons would not have been possible.

Meanwhile, other social science sub-disciplines likewise profit from their use of geospatial methods. The field of environmental justice, for

instance, gained such a massive attraction because social scientists were finally able to include environmental data in their analyses. Among many other findings, they discovered that deleterious environmental hazards affect members of ethnic minorities (Rüttenauer 2018), low-income groups (Zwickl, Ash, and Boyce 2014), or single parents (Downey, Crowder, and Kemp 2016) far more often than the general population. Environmental noise researchers showed that people with high scores of noise sensitivity suffer significantly more when they are exposed, e.g., to road traffic or air traffic noise (Stansfeld and Shipley 2015; Boes, Nüesch, and Stillman 2013). Accordingly, even more applications of geospatial methods can be found in an extensive collection of social sciences' sub-disciplines ranging from political behavior and attitudes (Dill and Jirjahn 2014; Förster 2018), educational (Ainsworth 2002; Crowder and South 2011; Weßling 2016) to health research (Bocquier et al. 2014; Saib et al. 2014; Oiamo et al. 2015).

With their use also come some challenges of georeferenced survey data. For example, researchers need to learn new methods and new concepts, e.g., by exploiting techniques that are only available in Geographic Information Systems (GIS). The following section introduces the challenges of particular small-scale geospatial data that, first of all, also affect privacy concerns.

Challenges of Georeferenced Survey Data

Usually, social science survey data contain only broad information on respondents' locations – if they contain location information at all. This information ranges from content information on municipality sizes or immigration rates to geographic information such as region or state names and other information on large-scale geographical boundaries. The reason for only providing this broad information on geographies is simple: in combination with other sociodemographic information, even relatively broad geographical information can impose a high risk of re-identifying single survey respondents and hence endanger data privacy (Skinner 2012; Duncan, Keller-McNulty, and Stokes 2003). Geographic information can create unique observations in datasets the more fine-grained this information is and the number of potential people with specific characteristics decreases. As an example, a lawyer with seven children for whom we know in which city she lives is easier to identify than a lawyer for whom we know in which country she lives. Therefore, data privacy concerns (Armstrong and Ruggles 2005) generally preclude the sharing of small-scale geographic information with other researchers and the public (Schweers et al. 2016), even if this information is available to the primary researchers.

To address the demand for analyzing survey data on the neighborhood level despite privacy concerns, data providers have implemented several distribution mechanisms over the past few years. Another possibility, not discussed here for the purpose of clarity, is to mask the data, either by

changing the geographic information (Allshouse et al. 2010; Zandbergen 2014) or other attributes of the data (Matloff and Tendick 2015). Often, a mixture of different approaches seems to be reasonable. Specialized contracts ensure that only researchers who have legal permission can access datasets, which are anonymized to not contain any personal information such as personal names, but are still sensitive. Alternatively, if the data are still too easy to de-anonymize, as for the case of the lawyer in a known city, research data centers provide on-site access facilities for limited and controlled access to the data (Goebel 2017) (https://www.gesis.org/en/services/data-analysis/more-data-to-analyze/secure-data-center-sdc/). With such an infrastructure in place, it is now increasingly possible to also access georeferenced survey data at the level of the respondents' addresses.

What makes georeferenced survey data so useful in social sciences research? Georeferenced survey data contain information based on direct spatial references. Direct spatial references identify locations not only using the names of, e.g., municipalities but also by identifiers such as geo-coordinates projected in a coordinate space. Moreover, in contrast with broad spatial information, these references often point to rather small-scale geographic information such as housing, street or neighborhood location.

Using GIS, researchers can project geographic information on survey respondents to analyze spatial patterns and to enrich their data with additional information (Meyer and Enzler 2013; Müller, Schweers, and Siegers 2017). Thus, within a GIS, each dataset embodies a single layer on a joint map. If we aim to enrich these data with additional information, the attributes of all other geospatial data are then systematically assigned to the corresponding information of the survey data. The result represents survey data that are structurally equivalent to the original survey data yet enriched with additional information from the other geospatial datasets.

We visualized this procedure using the example of the geo-coordinates of some fictional social science survey respondents combined with road traffic noise data, as shown in Figure 1. The figure shows a map section of the city of Cologne. On this map, a layer of road traffic noise is displayed, illustrating levels of objectively measured road traffic noise levels at different locations (German Environmental Agency/EIONET Central Data Repository 2016). The white points within this map are another layer showing the locations of fictional respondents' dwellings. Spatial linking procedures in a GIS then add the corresponding value of road traffic noise measurements to the geo-coordinates data of the fictional survey respondents.

In addition to the simple linking of layering feature attributes, other procedures are indeed possible. An often-used procedure is the calculation of geographic distances. Instead of assigning attributes based on features that are at the same location, the geographic distances (i.e., the euclidean distances) between a respondent's coordinate and the coordinate of a point

Figure 1 Map section of the city of Cologne with road traffic noise at its corresponding day, evening, and night mean decibel values and fictional respondents' geo-coordinates (data sources: OpenStreetMap Contributors 2017 and German Environmental Agency/EIONET Central Data Repository 2016).

of interest such as theaters, poll sites or hospitals can be calculated and added to the focal data location. At the same time, the resulting dataset does not differ in structure from the results of other spatial linking procedures: it consists of survey data identical to the original survey dataset, yet it is enriched with information such as geographic distances between specific geographic coordinates.

In principle, we can speak of georeferenced data as geospatial data as well. For the following reason, however, we refrain from using this denotation in this article. First, when researchers work with geospatial data, they often expect data in specific data formats, such as ESRI Shapefiles or GeoTiffs. Georeferenced survey data, however, are structured the same way as other ordinary survey data: in a rectangular data matrix stored as CSV, SPSS, Stata or other data format. Second, geospatial data represent information on geometries regardless of how small they might be. Georeferenced survey data, on the other hand, hold information on individuals scattered within a specific sample area that are at best only representative of all other individuals living in this area. Indeed, geospatial data might also be censored, but survey respondents are usually sampled based on probability. Therefore, georeferenced survey data constitute incomplete data by design, whereas geospatial data are usually targeted as being complete.

As noted, social science survey data linked to geospatial data attributes remain in the same rectangular data matrix format as the original survey data. However, the matrices provide additional information that is different from the information collected by interviewing respondents in a survey. For example, the process of collecting measurement data on environmental noise exposure is complicated. To make such data findable, replicable, and reusable, the information on it must be documented as thoroughly as for regular survey data. Hence, the documentation must not only be done at the content level but also concerns the collection of the geospatial data attributes and the method of linking them to the survey data.

Lastly, georeferenced survey data enriched with geospatial data attributes often does not contain direct spatial references. Storing survey information in combination with personal information, such as address-based geo-coordinates, on one single dataset is not allowed according to German data protection legislation (Müller 2019). For this purpose, these data are linked in multiple steps: first, the geo-coordinates are used to extract attributes from geospatial data; second, this information is added to the actual survey data by changing identifiers between the survey data and geo-coordinates and deleting the geo-coordinates (Schweers et al. 2016). This procedure is not the subject of this article – what remains important, however, is that even without direct spatial references, such as geo-coordinates, the linked data contain information stemming from geographic structures that must be described with appropriate metadata.

METADATA FOR GEOREFERENCED SOCIAL SCIENCE SURVEY DATA

Metadata for Social Science Survey Data and Geospatial Data

In the previous section, we presented a case in which we linked social science survey data with geospatial data. Although the final linked dataset does not differ in structure from the original dataset, the linked data required addressing specific documentation demands. In this section, we present different metadata standards and give reasons for the metadata standard we have chosen for our purpose, DDI.

In principle, there exists a manifold of different metadata standards for research data. To achieve the best documentation results possible, it is imperative to choose the most suitable metadata standard for one’s purposes. Some of the existing metadata standards are well known and used by many, whereas others are less known and possibly used by a single institution or project only. Various standards are discipline specific, whereas others are multidisciplinary. Furthermore, some metadata standards are very detailed and semantically rich; others contain only very few elements. Lastly, some standards even have the status of a norm.

An example of such a standard that is an ISO norm is the Dublin Core Metadata Element Set (DCMES 2012 – ISO Norm 15836). The Dublin Core Metadata Initiative began to develop the Dublin Core Metadata Standard in 1994. Its purpose has shifted from, originally, the documentation of literature, to the documentation of all digital objects that can be found on the internet, and, finally, to the documentation of all objects (digital or not) that can be identified. The Dublin Core Metadata Element Set contains 15 metadata elements which can be repeated (Jensen, Zenk-Möltgen, and Wasner 2019) (http://dublincore.org/). The Dublin Core standard is neither discipline specific nor very detailed or semantically rich; hence, it did not meet the needs for the documentation of the data of this article which are complex and cover different disciplines.

Another metadata standard we considered for documenting georeferenced survey data was the DataCite standard. DataCite is an international consortium established in 2009. It has members from Europe, North America, and Australia. The goal of DataCite is to identify research data and make it more visible and easier to locate. To achieve this, DataCite supports the creation and allocation of DOIs (Digital Object Identifiers, which are persistent identifiers based on the Handle system) and accompanying metadata. The DataCite metadata elements are metadata properties that accurately and consistently identify a resource for citation and retrieval purposes. The DataCite metadata schema contains six mandatory elements and 13 further elements which are optional for use (DataCite 2017; Jensen, Zenk-Möltgen, and Wasner 2019) (https://datacite.org/mission.html).

Both of the presented metadata standards contain elements for documenting geographic information, but specialized standards such as ISO 19115 provide possibilities to do that on a much more detailed level. For this reason, ISO 19115 is a widely used metadata standard for geospatial data worldwide. Additionally, the federal archives in Germany (AdV and KLA 2015) have integrated ISO 19115 into their data documentation. Furthermore, ISO 19115 strongly connects to the metadata elements used within the Infrastructure for Spatial Information in the European Community (INSPIRE) initiative, which will extend the availability of federal geospatial data in the coming years. If librarians rely heavily on open data distributed by federal agencies, using ISO 19115 for documentation is a cardinal choice.

Meanwhile, although other metadata standards for geographic information exist, georeferenced survey data are still the focal data of this article. These data differ from other geospatial data because they contain attributes collected from survey methods. Moreover, they often only contain information derived from other geospatial datasets; however, the actual location information (i.e., geo-coordinates) are either deleted or not easily accessible (see previous section). ISO 19115, for example, provides

tools to describe data from various disciplines, but the specifics of survey data cannot be covered by this standard. To our knowledge, this also applies to other standards for geospatial data.

Choice of DDI

As we use data from two different disciplines – where social science survey data are the focal data – our documentation must be very detailed and semantically rich but also discipline specific with regards to the whole dataset and each of its variables. Because the data originate in the use of a methodologically sophisticated data linking technique, the documentation must be detailed in such a way that others can subsequently reuse and comprehend the data. Additionally, the metadata should be machine actionable to facilitate its distribution to other digital data catalogs and to enhance the findability and visibility of the data. The limited number of metadata elements in each of the presented standards does not satisfy this purpose. At the same time, we are not aware of any previous attempts to integrate them for such a specific use.

A metadata standard that promises to meet these requirements, however, is the DDI Lifecycle Standard of the Data Documentation Initiative (DDI). DDI Lifecycle enables the extensive and thorough description of social science research data and their origin. DDI, in general, is the most elaborate and most commonly used metadata standard for social science survey data (Hoyle et al. 2011; Rasmussen 2014; Jensen, Zenk-Möltgen, and Wasner 2019). It is an international metadata standard that supports the documentation of data from the social, behavioral, economic, and health sciences, and it was initiated in 1995 as a project of the US-American archive ICPSR (Inter-University Consortium for Political Social Research). The aim was to improve the options for a standardized documentation of social science research data and their presentation on the internet, which is corroborated by the fact that many social science archives use this standard productively. DDI provides a detailed structure and facilitates machine-actionability and interoperability.

Currently two different DDI specifications exist, each having different versions of those specifications – DDI Codebook and DDI Lifecycle. DDI Codebook is less extensive than DDI Lifecycle and focuses on the after-the-fact documentation. The information included in DDI Codebook is on document description, study description, variable description and file description. DDI Lifecycle, on the other hand, offers more features than DDI Codebook. DDI Lifecycle allows the documentation of (mainly) social science research data across its life course. With DDI Lifecycle, all activities, from the conception of the study to the reuse of data, can be documented (see Figure 2).

Another advantage of DDI Lifecycle is that it has already integrated metadata fields for geospatial data (ISO 19115). Based on these

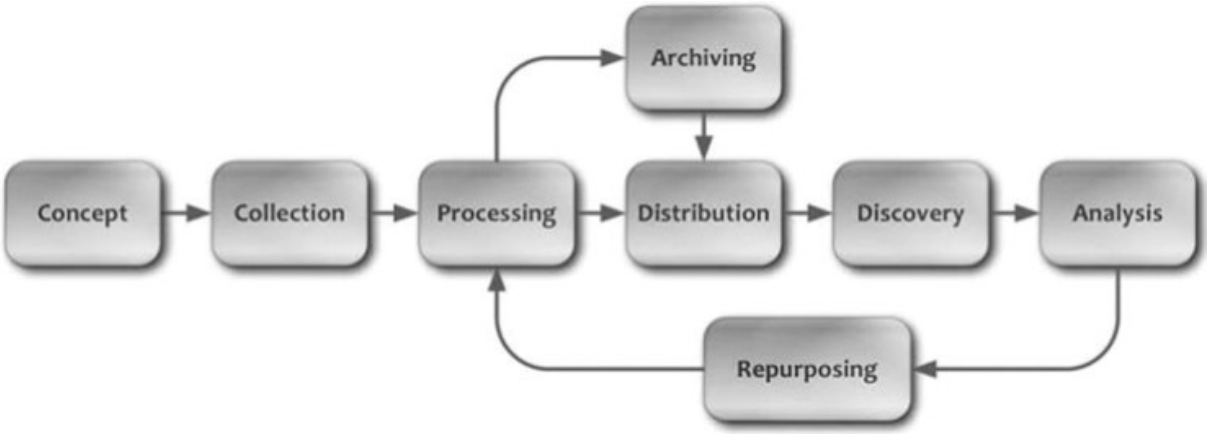

Figure 2 DDI Data Lifecycle. Source: http://www.ddialliance.org/training/why-use-ddi. © 2019 DDI Alliance. Reproduced by permission of DDI Alliance. Permission to reuse must be obtained from the rightsholder.

considerations, we determined DDI to be the most natural choice for documenting georeferenced survey data (Green and Humphrey 2013; Jensen, Zenk-Möltgen, and Wasner 2019; Hoyle et al. 2011; Rasmussen 2014; Zenk-Möltgen 2012; Vardigan, Heus, and Thomas 2008):

> *"With respect to the geographic standards, DDI developers consulted with geographers and experts in geospatial data to ensure that the DDI captures the core elements needed for resource discovery of social science data without pulling in the bulk of these larger standards."* (Vardigan, Heus, and Thomas 2008, *110*).

The Current State of ISO 19115 in DDI Lifecycle

In general, there is indeed a correspondence between DDI and ISO 19115. Within DDI, we can draw on capabilities to describe space and its attributes for which the creators of the DDI standard defined specific ISO 19115 elements. For example, metadata fields of ISO 19115 used to describe features such as geographic structures, bounding boxes, or coordinate reference systems already map to corresponding DDI metadata fields. A collection of mappings is shown in Table 1.

At the same time, only a restricted number of the ISO 19115 terms were included in DDI as the idea within DDI is not to incorporate all metadata fields of the ISO 19115 standard but to enable a link to a geospatial raster or vector file with more information. ISO 19115 metadata on geographic structures can be integrated into DDI at the study level – the level of data that describes a dataset as a whole. This way, we can describe geographic structures for all variables of a dataset at once, but we cannot define different geographic structures for each variable separately. While ISO 19115 metadata can be used in DDI also for variables that contain

TABLE 1 Mapping of ISO 19115 Terms to DDI Lifecycle Terms

ISO 19115 term	DDI Lifecycle term
bounding	BoundingBox
westbc	WestLongitude
eastbc	EastLongitude
northbc	NorthLatitude
southbc	SouthLatitude
dsgpolyo	BoundingPolygon
dsgpolyx	ExcludingPolygon
timeperd	GeographicTime

geographic information as values, (e.g. country codes, bounding boxes or geo-coordinates), this is not possible for variables which comprise contents with underlying geographic structures. For example, DDI does not provide a reference geographic structure of variables such as unemployment rates at the district or city level. This missing feature will not create a problem if all variables derive from one single geospatial dataset. However, we assume that this is not a particularly realistic scenario, and it certainly is not the case for our data as we show in the subsequent sections.

Our example of the georeferenced GGSS 2014, which is elaborated even further in our case study below, clarifies this point. After we georeferenced the survey data, we initially linked them to environmental noise data and subsequently to data from the German Census 2011. These data stem from different sources. They exist not only in different formats but also in different geographic structures. Overall, we faced a heterogeneous set of new attributes that we could not describe with the standard implementation of ISO 19115 in DDI because of its limitations described earlier.

Currently, no solutions are available for these documentation issues; hence, we developed four workarounds that facilitate the documentation of georeferenced survey data at the variable level. The workarounds differ in the extent to which they are still valid with regards to the original specification. In the following section, we describe the implementations and the consequences of those approaches before we demonstrate the necessity of these considerations in our case study.

APPROACHES TO NAVIGATING DOCUMENTATION CHALLENGES

Workaround Approaches Respecting DDI Validity

The first and the second workarounds have some convenient features. Despite not complying with the original implementation of a standard, they do not break the standard in the sense of producing incompatibilities. Resulting documents are still formally valid, even when there are limitations in the semantic meaning of the content of the elements. Furthermore,

these workarounds are probably easy to implement because they can be used in accordance with existing systems and software (https://www.ddialliance.org/resources/tools). However, in the long run, as datasets and corresponding metadata grow, these workarounds might become difficult to maintain. Here, we present two workarounds for documenting georeferenced survey data, each of which has advantages as well as disadvantages.

Our **first workaround** consists of splitting the dataset logically into different studies, for which each has its own DDI metadata (using the DDI element *StudyUnit*). Each of the subsets of the original dataset comprises only attributes from geospatial data that stem from one single data source. Accordingly, this procedure assigns individual metadata objects to each of these subsets. For example, if we link three different geospatial datasets to one single georeferenced survey dataset, we receive three distinct studies. This workaround adds complexity to the management of the dataset because it originates from several studies.

When using this workaround, metadata on geographic structures for each study refer to each study as a whole. However, because each study only contains information from one single geospatial dataset with one single geographic structure, geographic structures can be described appropriately. This information can be linked to any variable in each dataset as well as across studies using the same reference. Hence, by applying this approach, georeferenced survey data linked to geospatial information can be accurately documented.

The most significant advantage of this approach is that DDI as the standard for documenting the data is still in use. The documentation is compatible with existing systems, for example, systems that can catalog the data and foster the reuse of metadata and data. At the same time, data librarians will end up with an increased amount of separate metadata objects for one spatial linking project. In addition, this procedure lacks a description for the combined dataset, which consists of all the linked data sources. For these reasons, we present an alternative workaround that might be more appropriate for spatial linking projects that use data from several different data sources.

This **second workaround** involves not using DDI at all for the description of the geospatial data attributes. In this case, separate files with structured metadata are created, (e.g., ISO 19115 metadata), that can again be referenced in DDI at the variable level. Consequently, the initial DDI metadata object that describes a single social science study linked to geospatial data attributes remains a single entity. Only the geospatial data attributes remain separate entities, as they stem from different datasets. Because data librarians can rely on standards such as ISO 19115, they can also draw on the whole set of metadata to describe geospatial attributes instead of just a small subset implemented in DDI.

An improved variant of this workaround is to include the ISO 19115 metadata for a specific variable in a DDI *Note* element. This variant enables to maintain DDI-compliant documentation that has the complete information available; it does not require managing additional files for the geospatial attributes. However, the management of the ISO 19115 metadata elements is still needed to maintain the documentation.

Again, the most significant advantage of this approach is that DDI as the standard for documenting the survey data is still in use, and the DDI file remains valid. The most significant disadvantage, however, is that the number of separated metadata objects increases as the amount of information from different geospatial data attributes increases, and this holds true for both workarounds as shown above. This disadvantage can be avoided by using the described variant of the second workaround and by including the attributes within a *Note* element for each variable. However, even then this procedure still lacks the semantic integration of the attributes into the DDI standard. Furthermore, both workarounds indeed do not break the DDI standard, but in the case of the second workaround, data librarians also operate the geospatial information independent of DDI. Given the advantages as well as disadvantages, it is up to individual projects to decide which workaround would comply best with the project's demands.

Workaround Approaches Ignoring DDI Validity

There is still another set of useful approaches to the documentation issue: workarounds that break the standard so that validation with the original specification is no longer possible. In general, ignoring validity means implementing documentation features that the standard does not provide. Thus, it forces the implementation of features that data librarians require to see achieved. Choosing this approach breaks the standard and causes incompatibilities with other systems. Users of such an approach therefore develop their own system that can be used for their specific application. Related to DDI, we again present two distinct approaches.

For the **third workaround**, data librarians can use the existing options for describing geographic structures in DDI at the study level and apply them to the variable level. Because this is not supposed to be done according to the DDI specification, it would result in metadata files that are incompatible with the official implementation of DDI. At the same time, DDI metadata objects are simple text data, stored in an XML format. Therefore, XML elements can indeed be injected into the XML structure at undesignated places. By doing this, simple routines can be developed that store DDI geographic structures in a structured way.

For the **fourth workaround**, data librarians can use the existing option of describing geographic structures in DDI as designed at the study level, but use that option multiple times for different geographic structures. Because using geographic structure references for variables in DDI is not allowed, this would break the formal implementation design. However, this approach is less invasive than the third workaround approach. It operates at the study level and only inserts references at the variable level that share the defined geographic structure. Furthermore, implementing this approach would be easier to accomplish because it does not require finding a proper position within the whole XML structure of DDI to inject geographic structure fields. They are simply written in the XML in succession.

Choosing the third or fourth workaround comes at a price. Both approaches would render the XML document invalid to DDI. If a project works in a setting that relies heavily on the general implementation of the DDI standard, the connection to other cataloging systems will be adversely affected. An example would be the connection to data registration systems that access a cataloging system. If the documentation breaks the standard, these systems could not process the data they receive. If exchanging metadata with other DDI-based systems is not intended, however, this disadvantage might be less serious. In this case, data librarians could also use an adapted XSD (XML Schema Document) of DDI to validate their XML files.

Similar to the first workaround, a striking advantage of both the third and fourth workaround approaches is that they can be applied entirely within DDI. Moreover, as the number of linked geospatial data attributes increases, the number of separated files or referenced metadata objects does not increase at all. When working on complex spatial data linking projects with an increased number of different data sources, the complexity of documenting them would be kept low. There is no need to create a set of intertwined metadata objects – one single metadata object is sufficient.

The following section further demonstrates the necessity of applying one of these workarounds. We present a case study of the Georeferenced German General Social Survey 2014 that was enriched with geospatial information from road traffic noise measurements and immigrant rates. After introducing the data, we determine in detail the contrast between what we would like to describe from this information and what is possible to describe with the actual metadata implementation.

CASE STUDY: GEOREFERENCED GENERAL SOCIAL SURVEY 2014

Data and Spatial Linking

Our focal source of social science survey data is the German General Social Survey 2014 (GGSS) (GESIS - Leibniz Institute for the Social Sciences

2015, 2018). The GGSS is a sample comprising private households in Germany and respondents who are at least 18 years old at the time of the interview. This survey is conducted every two years with the aim of monitoring trends in attitudes, behavior, and societal change in the Federal Republic of Germany. The GGSS is a well-known dataset that is often used by researchers from Germany and abroad (GESIS-Data Archive For The Social Sciences 2018).

We used two sources of geospatial data to link to the survey data of the GGSS. The first data source was geospatial data collected in correspondence with the Environmental Noise Directive (2002/49/EC) of the European Union (EU) (European Parliament and European Council 2002). This directive obligates members of the EU to collect data on noise originating from industrial facilities as well as from air, rail, and road traffic. While acquiring these data for Germany is challenging because of availability issues (Schweers et al. 2016), in principle, they consist of categorized polygon data that capture noise pressure measurements for each noise source. For the case study, we used the subset of road traffic noise data (German Environmental Agency/EIONET Central Data Repository 2016). The second data source was geospatial data that were collected in 2011 within the 2008 European Union census regulation (763/2008) (European Parliament and European Council 2008). These data aim to monitor demographic compositions of the population on a small scale. For Germany, the data are available on 1 km² aggregated grid cell attributes that extend to the whole area of Germany of which we used the information on immigrant rates (Statistical Offices of the Federation and the Länder 2018). Thus, the geospatial data we linked to the survey data of the GGSS consisted of a distinct set of datasets regarding the data format and content, which resulted in a significant impact on the metadata, as we describe below.

In compliance with German data protection legislation, we geocoded the GGSS respondents' addresses, assigned attributes of geospatial data and linked these attributes to the survey data attributes of the GGSS. Conceptual and technical background information can be found in (Schweers et al. 2016). While this process involves applying complex GIS methods of spatial linking such as building buffers or calculating geodesic distances, for the purpose of this case study we solely present the results of spatial linking by location. This method uses the geo-coordinates of the survey respondents, projects them in one coordinate space with the geospatial data, and extracts attributes of the latter. As a result, we gathered information on dB(A) road traffic noise values and immigrant rates as percentages for each survey respondent.

The structure of the resulting data is not different from that of ordinary survey data. Table 2 shows a fictional example of the GGSS with additional information on road traffic noise and immigrant rates. As we can

TABLE 2 Structure of the Survey Data Enriched with Road Traffic Noise Measurements and Immigrant Rates

ID	Survey Question 1	…	Survey Question k	Road Traffic Noise	Immigrant Rates
1	5	…	“maybe”	55	8.90
…	…	…	…	…	…
n	2	…	“yes”	75	34.78

see, the dataset is arranged in a spreadsheet format containing observations (i.e., the survey respondents) in the rows and attributes (i.e., survey answers and geospatial information) in the columns. Accordingly, all efforts of adding new geospatial information to georeferenced survey data result in a common survey data format with some additional columns that are the geospatial information.

Documenting the Linked Data

The resulting dataset of our case study – as shown above – consisted of variables resulting from the survey data, variables from the noise dataset, and variables about immigrant rates. Documenting all of the variables in the dataset with corresponding metadata was necessary for secondary researchers to work with the data. However, doing this in an appropriate way by using the DDI Lifecycle standard had some limitations, as described above and further explained in the following. For this purpose, we used two variables from the dataset: *roadm* “Road Traffic Noise” and *immi* “Immigrant Rates” (see Table 3).

First, we searched for specific DDI Lifecycle elements that could be used for documenting the geospatial metadata that we had. Because we were documenting variables, this needed to be elements within or referenced by the DDI Lifecycle element *Variable*. All variables of the dataset within the study were located at the XPath/*ddi:DDIInstance/s:StudyUnit/l:LogicalProduct/l:VariableScheme/l:Variable*. For seven of the metadata fields, we could find DDI Lifecycle elements within the *Variable* element: *VariableName*, *Label*, *Description*, *SourceUnit*, *MeasurementUnit*, *VariableRepresentation*, and *CodeListReference* (see Table 3). For other metadata, (e.g., the year, data type, geometry type, geographic extent, or coordinate reference system), we could not find appropriate elements within DDI Lifecycle for variable documentation. One possible solution to this was to aggregate all other metadata into the *Description* element, but this would have resulted in unstructured documentation, which was not our intention.

In a second step, we examined the *GeographicStructure* element within DDI Lifecycle that documents geographic metadata according to the

TABLE 3 Metadata Fields for Example Variables of Road Traffic Noise and Immigrant Rates

Metadata Field	Example Variable I	Example Variable II	DDI-L Element
Variable name	roadm	immi	l:VariableName
Label	Road Traffic Noise	Immigrant Rates	r:Label
Description	Road traffic noise at main roads, day-evening-night-mean	Rate of people without German citizenship	r:Description
Source	German Environmental Agency / EIONET Central Data Repository	Statistical Offices of the Federation and the Länder	l:SourceUnit
Year	2012	2011	
Measurement unit	dB(A)	%	l:VariableRepresentation/ r:MeasurementUnit
Scale	categorical	continious	l:VariableRepresentation
Value range	50-54, 55-59, 60-64, 65-69, 70-74, 75+	0-100	l:VariableRepresentation/ r:CodeRepresentation/ r:CodeListReference
Censored?	below 50 dB(A)	below 3 people on which percentages are based	
Data type	vector data	raster data / CSV	
Geometry type	polygons	raster cells	
Geographic extent of geometries	varying, changes within meters	1 km², uniformly shaped	
Coordinate Reference system	varying, harmonized in EPSG:3035	EPSG:3035	
Spatial Linking Sensitivity (SLS)	high, placement of geo-coordinates matters	low, placement of geo-coordinates matters only at borders	
Errors because of SLS	very likely	unlikely	

DDI specification. Within the *StudyUnit* element of DDI Lifecycle there is usually a reference to one or more *GeographicStructure* and to *GeographicLocations*. The *GeographicStructure* describes different *GeographicLevels*, (e.g., Country, Province, and Municipality). The *GeographicLocation* defines their specific instances, e.g., Germany (with a reference to the *GeographicLevel*). By using this approach, we could define the coverage of a study to be Germany. We could now start using the same possibilities within DDI Lifecycle to document the two example variables (see Figure 3). For variable *roadm* this would include defining a *GeographicStructure* with several *GeographicLevels* for the different noise categories (e.g., in 5 dB(A), in 3 dB(A), or in 1 dB(A) groups) and link the variable to the appropriate one (the 5 dB(A) level).

Figure 3 Usage of DDI Lifecycle *GeographicStructures* and *GeographicLocations*.

Unfortunately, a link from the *Variable* element to a *GeographicLevel* was not possible. For the second example variable *immi*, we had the same situation: *GeographicLevels* could be defined for different grid cells (e.g., 10.000 km², 100 km², and 1 km²), but a link to the appropriate 1 km² level was not possible (see Figure 3). What was possible within the dataset was having a variable for the location of the respondents that held the information about the exact grid cell of the person. This could use values from a *GeographicLocation* (e.g., saying that the grid instance D7 is where the respondent lives). Via the *GeographicLocation* we would also be able to link to the respective *GeographicLevel*. However, this solution was not possible in our case because of the privacy restrictions: We did not want to include the location of the respondents in the dataset; we want to keep only the road traffic noise level and immigrant rates.

This case study showed in more detail that the possibilities of DDI Lifecycle were limited with regards to the documentation of the actual linked dataset. As such, data librarians and research projects must rely on one of the workarounds described earlier. What remains is the actual choice for a workaround, which can differ depending on the goals of

individual projects. In the following Discussion we provide a general scheme that aims to help in choosing between the workarounds.

DISCUSSION

Which approach should projects use to document their georeferenced survey data linked to geospatial data attributes? The answer to this question depends on the documentation requirements that are considered important within each individual project. As the existing implementation of ISO 19115 in DDI Lifecycle is insufficient for many purposes, navigating this challenge means aiming for the best results given the requirements and possible disadvantages. This means that for varying use cases, varying approaches are best suited. Thus, there is no final answer to the question of the best approach.

However, we propose general criteria that can be helpful in choosing an individual approach:

- Do the metadata have to be valid DDI instances?
- How many datasets are involved? How many social science survey datasets are there? How many geospatial datasets are there?
- How many actors and stakeholders are involved in the processing of the data and metadata?
- Are the metadata exchanged with other actors using the DDI standard? What is the role of DDI in the institution, and is compliance with an existing standard important?

Depending on the answers to those questions, the choices will differ. If researchers, projects, and data librarians work in a closed environment and they do not exchange data or metadata, the choice of the approach does not matter. In this case, one of the third or fourth workarounds will probably be the best approach because these approaches contain all of the necessary information within one single object. However, this is a rare and unlikely scenario. For example, if researchers intend to give their data and metadata to a data archive working with DDI to engage cataloging that will be open for harvesting, invalid DDI documents will be ineligible. For these reasons, researchers must make a weighted decision between several different options.

Table 4 shows how specific conditions regarding the criteria result in different choices. Each column depicts the approaches described above to documenting georeferenced survey data: workarounds 1 & 2 as well as workarounds 3 & 4. Each row displays the general criteria. The cell combinations, therefore, represent the impact of each approach on each general

TABLE 4 Criteria that Lead to a Selection Between the Workarounds Respecting and Ignoring DDI Validity

Criterion		Workarounds Respecting DDI Validity		Workarounds Ignoring DDI Validity	
		1 Split into studies	2 Use ISO files for variables	3 Geospatial DDI for variables	4 Geospatial DDI referenced by variables
Need valid DDI	Yes				
	No				
Number of data collections	Many				
	Few				
Number of actors involved	Many				
	Few				
Exchange of DDI	Yes				
	No				
Standard compliance	Important				
	Unimportant				

criterion, signaled either by positive (green) or negative (red) indicators. Depending on the results, researchers, projects, and data librarians then can evaluate as well as weight the corresponding consequences and decide if these consequences are tenable. For example, if a project relies on valid DDI documents, both the third and fourth workaround approaches are ineligible because they would most certainly result in invalid DDI documents.

In fact, our example of linking geospatial data attributes to the survey data of the GGSS was one such example, for which we required a valid DDI XML document. Because the GESIS Data Catalog (https://dbk.gesis.org/dbksearch/) connects to other agencies that all use the official DDI specification, violating the standard would have imposed incompatibilities. Hence, workarounds three or four were not an option. Because the GGSS 2014 is an extensive study, using the first workaround (splitting the dataset logically into different studies, for which each has its own DDI metadata document) was not a preferred approach either; it would have been too much work, and the resulting number of different study descriptions would have become confusing. Therefore, in our use case, we preferred to use the second workaround and not describe the geospatial data in DDI, but rather in separate (non-DDI) files, which we referenced within our DDI file. Moreover, to avoid having to manage many separate files, we preferred the variant of including the geospatial metadata into the mentioned *Note* element of the variables.

CONCLUSION

How can researchers, projects and data librarians document georeferenced survey data that are linked to multiple sources of geospatial data? Generally, because both data types already provide subject-specific metadata standards, DDI Lifecycle and ISO 19115, it is possible to document both types of data in these standards. In this article, however, our focus was to describe them after they were linked. The result was that one can, in fact, document geospatial data attributes to a certain extent in DDI Lifecycle, but not to the extent necessary for the data. For these reasons, we developed four approaches to documentation that were suited to a broad range of different projects:

The first workaround was to logically split the dataset into different studies. Each of those different studies was described in its own DDI metadata document. Using this approach, the geo-information referred to the level of data it needed to refer to and the DDI document also remained valid. The second workaround was that only one DDI document existed for the study, and the information that cannot be captured within DDI is described in the ISO 19115 metadata standard, which provides the necessary possibilities. The non-DDI metadata could either be stored in separate files and be referenced from the main DDI metadata document or could be

included into a *Note* element within the DDI document. In any case, the DDI document would also be valid with this workaround. For the third workaround, the existing elements to describe geographic structures in DDI at the *variable level* are used, which is not valid in the design of DDI. As a fourth workaround approach, again the existing options of DDI are used; however, they are used at the *study level* and multiple times for the different geographic structures. Both the third and fourth workaround approaches would render the DDI document invalid.

For our case of the GGSS 2014, the third and fourth workarounds were not an option because we needed a valid DDI document. We preferred the second workaround with the variant, where the social science survey data are described in DDI and the geospatial metadata in ISO 19115 are included within a *Note* element of the DDI metadata. This ensured that the DDI documents were valid and, therefore, were compatible with existing data cataloging and data registration systems.

In general, applying workarounds is not a bad option. The use of workarounds demonstrates that metadata standards are flexible and indeed utilizable. It would be better, however, if all of the documentation needed could be contained in one file and, at the same time, we could have a valid metadata document. Consequently, future projects should focus on standardizing approaches to documenting these data, especially in the context of developing machine-actionable approaches.

As we have shown, possibilities exist to capture metadata information on geospatial data linked to social science data using the DDI Lifecycle metadata standard. However, those possibilities are not ideal and further work needs to be done. As the operational phase of the Infrastructure for Spatial Information in the European Community (INSPIRE) initiative is approaching (https://inspire.ec.europa.eu/inspire-roadmap/), the amount of available geospatial data that can be linked to social survey data will increase massively in the coming years. The same holds true for other types of data, such as social media or experimental data, the use of which is becoming more common in social science research. Providing standardized ways of documenting these non-survey data types, especially when they are linked to other datasets, is therefore one of the most vital and possibly most challenging tasks for the data documentation community in the coming years.

Projects such as ours are necessary to provide use cases for the developers of metadata standards and, if possible, to support metadata initiatives, e.g., by delivering content-related input on missing metadata elements. The purpose of such work can vary; depending on the requirements, the ideal could even be to develop a whole new standard or just to extend existing ones. As most metadata standards are a product of years of development in a particular area, discarding all of these efforts may not be

the best approach. Standards were developed for a reason, and for most applications they should, in principle, apply. Thus, we argue that projects and developers should first always consider integrating their work into existing standards.

Moreover, actively communicating with the metadata initiatives and their developers would be the second step. We have done this twice, by presenting and discussing our progress in documenting georeferenced survey data (Müller, Schweers and Zenk-Möltgen 2015, 2016) and by writing this article. We hope that we have taken a first step towards standardization in documenting georeferenced survey data. In our case, which we believe has similar needs as other social science studies with georeferenced survey data, it may become possible to document geo-information at the variable level for each variable separately in the future. Therefore, we need other projects that face similar problems to step forward and showcase their issues. In this way, metadata standard initiatives can develop ideas on how to extend their standards so that the use of workarounds, presented here, becomes obsolete. At the same time, data librarians can continue to rely on the metadata standards they are most familiar with and which provide them with the tools to accurately document their data, which in our case is the DDI standard.

ORCID

Stefan Jünger http://orcid.org/0000-0001-8100-7957
Kerrin Borschewski http://orcid.org/0000-0002-0284-7024
Wolfgang Zenk-Möltgen http://orcid.org/0000-0002-2158-3941

REFERENCES

AdV and KLA. 2015. “Leitlinien Zur Bundesweit Einheitlichen Archivierung von Geobasisdaten. Abschlussbericht Der Gemeinsamen AdV-KLA-Arbeitsgruppe ‘Archivierung von Geobasisdaten’ 2014-2015. Hamburg. http://www.bunde-sarchiv.de/DE/Content/Downloads/KLA/leitlinien-geobasisdaten.pdf.

Ahonen-Rainio, Paula. 2006. Metadata for geographic information. *Journal of Map & Geography Libraries* 2 (1):37–66. doi: 10.1300/J230v02n01_03.

Ainsworth, James W. 2002. Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces* 81 (1):117–52. doi: 10.1353/sof.2002.0038.

Allport, Gordon W. 1954. *The nature of prejudice*. Cambridge, MA: Addison-Wesley Publishing Company.

Allshouse, William B., Molly K. Fitch, Kristen H. Hampton, Dionne C. Gesink, Irene A. Doherty, Peter A. Leone, Marc L. Serre, and William C. Miller. 2010. Geomasking sensitive health data and privacy protection: An evaluation using an E911 database. *Geocarto International* 25 (6):443–52. doi: 10.1080/10106049.2010.496496.

Armstrong, Marc P., and Amy J. Ruggles. 2005. Geographic information technologies and personal privacy. *Cartographica: The International Journal for Geographic Information and Geovisualization* 40 (4):63–73. doi: 10.3138/RU65-81R3-0W75-8V21.

Blalock, Hubert M. 1967. *Toward a theory of minority-group relations*. New York: Wiley.

Bluemke, Matthias, Bernd Resch, Clemens Lechner, René Westerholt, and Jan-Philipp Kolb. 2017. Integrating geographic information into survey research: Current applications, challenges and future avenues. *Survey Research Methods* 11 (3):307–27. https://doi.org/10.18148/srm/2017.v11i3.6733.

Blumer, Herbert. 1958. Race prejudice as a sense of group position. The Pacific Sociological Review 1 (1):3–7. doi: 10.2307/1388607.

Bocquier, Aurélie, Sébastien Cortaredona, Céline Boutin, Aude David, Alexis Bigot, Basile Chaix, Jean Gaudart, and Pierre Verger. 2014. Is exposure to night-time traffic noise a risk factor for purchase of anxiolytic-hypnotic medication? A cohort study. European Journal of Public Health 24 (2):298–303. doi: 10.1093/eurpub/ckt117.

Boes, Stefan, Stephan Nüesch, and Steven Stillman. 2013. Aircraft noise, health, and residential sorting: Evidence from two quasi-experiments: Aircraft noise and health. *Health Economics* 22 (9):1037–51. doi: 10.1002/hec.2948.

Crowder, Kyle, and Scott J. South. 2011. Spatial and temporal dimensions of neighborhood effects on high school graduation. *Social Science Research* 40 (1): 87–106. doi: 10.1016/j.ssresearch.2010.04.013.

DataCite. 2017. DataCite Metadata Scheme 4.1. Accessed February 15, 2019 https://schema.datacite.org/meta/kernel-4.1/.

Dill, Verena, and Uwe Jirjahn. 2014. Ethnic residential segregation and immigrants' perceptions of discrimination in West Germany. *Urban Studies* 51 (16): 3330–47. doi: 10.1177/0042098014522719.

Downey, Liam, Kyle Crowder, and Robert J. Kemp. 2016. Family structure, residential mobility, and environmental inequality: Family structure and environmental inequality. *Journal of Marriage and Family* 79 (2):535–55. doi: 10.1111/jomf.12355.

Duncan, George T., Sallie A. Keller-McNulty, and S. Lynne Stokes. 2003. Disclosure risk vs. data utility: The R-U confidentiality map. *CHANCE* 17 (3): 16–20.

Edwards, Paul N., Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41 (5):667–90. doi: 10.1177/0306312711413314.

European Parliament and European Council. 2002. "Directive 2002/49/EC of the European Parliament and of the European Council." http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32002L0049.

European Parliament and European Council. 2008. REGULATION (EC) No 763/2008 on Population and Housing Censuses.

Förster, André. 2018. Ethnic heterogeneity and electoral turnout: Evidence from linking neighbourhood data with individual voter data. *Electoral Studies* 53 (June):57–65. https://doi.org/10.1016/j.electstud.2018.03.002.

German Environmental Agency/EIONET Central Data Repository. 2016. “Road Traffic Noise 2012 Shapefiles.” Accessed November 30, 2016, https://github.com/stefmue/georefum/blob/master/data/cdr.road.lden.dat.rda.

GESIS - Leibniz Institute for the Social Sciences. 2015. “ALLBUS/GGSS 2014 (Allgemeine Bevölkerungsumfrage Der Sozialwissenschaften/German General Social Survey 2014). *GESIS Data Archive*. http://dx.doi.org/10.4232/1.12209.

GESIS - Leibniz Institute for the Social Sciences. 2018. “ALLBUS/GGSS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey) - Sensitive Regional Data.” GESIS Data Archive. https://doi.org/10.4232/1.13010.

GESIS-Data Archive For The Social Sciences. 2018. “Downloadstatistik GESIS Datenarchiv.” GESIS Data Archive. https://doi.org/10.4232/1.12979.

Goebel. 2017, Jan. SOEP 2015-Informationen Zu Den SOEP-Geocodes in SOEP V32. SOEP Survey Papers 407. DIW.

Gómez, Nancy-Diana, Eva Méndez, and Tony Hernéndez-Pérez. 2016. Data and metadata research in the social sciences and humanities: An approach from data repositories in these disciplines. *El Profesional de La Información* 25 (4): 545. doi: 10.3145/epi.2016.jul.04.

Green, Ann, and Chuck Humphrey. 2013. Building the DDI. *IASSIST Quarterly* 37 (1–4):36–44. doi: 10.29173/iq500.

Hillmert, Steffen, Andreas Hartung, and Katarina Weßling. 2017. Dealing with space and place in standard survey data. Survey Research Methods, Vol 11, No 3 (2017): Special Issue: Uses of Geographic Information Systems Tools in Survey Data Collection and Analysis, October. https://doi.org/10.18148/srm/2017.v11i3.6729.

Hoyle, Larry, Fortunato Castillo, Benjamin Clark, Neeraj Kumar Kashyap, Denise Perpich, Joachim Wackerow, and Knut Wenzig. 2011. Metadata for the longitudinal data life cycle: The role and benefit of metadata management and reuse. DDI Alliance Working Papers Series doi: 10.3886/DDILongitudinal03.

Jensen, Uwe, Alexia Katsanidou, and Wolfgang Zenk-Möltgen. 2011. Metadaten und Standards. In *Handbuch Forschungsdatenmanagement*, eds. Stephan Büttner, Hans-Christoph Hobohm, and Lars Müller. Bad Honnef: Bock + Herchen Verlag.

Jensen, Uwe, Wolfgang Zenk-Möltgen, and Catharina Wasner. 2019. Metadatenstandards Im Kontext Sozialwissenschaftlicher Daten. In *Forschungsdatenmanagement Sozialwissenschaftlicher Umfragedaten. Grundlagen Und Praktische Lösungen Für Den Umgang Mit Quantitativen Forschungsdaten*, eds. Uwe Jensen, Sebastian Netscher, and K atrin Weller, 151–78. Opladen, Berlin, Toronto: Verlag Barbara Budrich.

Jünger, Stefan. 2019. Using georeferenced data in social science survey research: The method of spatial linking and its application with the German General Social Survey and the GESIS Panel (GESIS Series, 24). Cologne: GESIS Leibniz Institute for the Social Sciences. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-63688-7

Klinger, Julia, Stefan Müller, and Merlin Schaeffer. 2017. Der Halo-Effekt in Einheimisch-Homogenen Nachbarschaften: Steigert Die Ethnische Diversität

Angrenzender Nachbarschaften Die Xenophobie? *Zeitschrift Für Soziologie* 46 (6):402–19. doi: 10.1515/zfsoz-2017-1022.

Kong, Nicole Ningning. 2015. Exploring best management practices for geospatial data in academic libraries. *Journal of Map & Geography Libraries* 11 (2): 207–25. doi: 10.1080/15420353.2015.1043170.

Legewie, Joscha, and Merlin Schaeffer. 2016. Contested boundaries: Explaining where ethnoracial diversity provokes neighborhood conflict. *American Journal of Sociology* 122 (1):125–61. doi: 10.1086/686942.

Martig, Noemi, and Julian Bernauer. 2016. Der Halo-Effekt: Diffuses Bedrohungsempfinden Und SVP-Wähleranteil. *Swiss Political Science Review* 22 (3):385–408. doi: 10.1111/spsr.12217.

Matloff, Norman, and Patrick Tendick. 2015. A new method for avoiding data disclosure while automatically preserving multivariate relations. arXiv: 1510.04406.

Meyer, Reto, and Heidi Bruderer Enzler. 2013. Geographic Information System (GIS) and its application in the social sciences using the example of the Swiss environmental survey. *MDA* 7 317–346. doi: 10.12758/mda.2013.016.

Müller, Stefan, Stefan Schweers, and Wolfgang Zenk-Möltgen. 2015. Georeferenced survey data at the GESIS data archive. Presented at the EDDI15 – 7th annual European DDI user conference, Copenhagen.

Müller, Stefan, Stefan Schweers, and Wolfgang Zenk-Möltgen. 2016. The past, present and future of geocoded survey data at the GESIS data archive. Presented at the EDDI16 – 8th annual European DDI user conference, Cologne.

Müller, Stefan. 2019. Räumliche Verknüpfung Georeferenzierter Umfragedaten Mit Geodaten: Chancen, Herausforderungen Und Praktische Empfehlungen. In *Forschungsdatenmanagement Sozialwissenschaftlicher Umfragedaten. Grundlagen Und Praktische Lösungen Für Den Umgang Mit Quantitativen Forschungsdaten*, eds. Uwe Jensen, Sebastian Netscher, and Katrin Weller, 211–29. Opladen, Berlin, Toronto: Verlag Barbara Budrich.

Müller, Stefan, Stefan Schweers, and Pascal Siegers. 2017. Geocoding and spatial linking of survey data – An introduction for social scientists. *GESIS Paper* 15: 1–29.

Nonnenmacher, Alexandra, and Juergen Friedrichs. 2011. The missing link: Deficits of country-level studies. A review of 22 articles explaining life satisfaction. *Social Indicators Research* 110 (February):1221–44. doi: 10.1007/s11205-011-9981-8.

Oiamo, Tor H., Jamie Baxter, Alice Grgicak-Mannion, Xiaohong Xu, and Isaac N. Luginaah. 2015. Place effects on noise annoyance: cumulative exposures, odour annoyance and noise sensitivity as mediators of environmental context. *Atmospheric Environment* 116 (September):183–93. doi: 10.1016/j.atmosenv.2015.06.024.

OpenStreetMap Contributors. 2017. Planet dump. https://Planet.Osm.Org.

Porcal-Gonzalo, Maria C. 2015. A strategy for the management, preservation, and reutilization of geographical information based on the lifecycle of geospatial data: An assessment and a proposal based on experiences from Spain and Europe. *Journal of Map & Geography Libraries* 11 (3):289–329. doi: 10.1080/15420353.2015.1064054.

Rasmussen, Karsten Boye. 2014. Social Science Metadata and the Foundations of the DDI. *IASSIST Quarterly* 37 (1):28. doi: 10.29173/iq499.

Rüttenauer, Tobias. 2018. Neighbours matter: A nation-wide small-area assessment of environmental inequality in Germany. *Social Science Research* 70 (February):198–211. doi: 10.1016/j.ssresearch.2017.11.009.

Rydgren, Jens, and Patrick Ruth. 2013. Contextual explanations of radical right-wing support in Sweden: Socioeconomic marginalization, group threat, and the halo effect. *Ethnic and Racial Studies* 36 (4):711–28. doi: 10.1080/01419870.2011.623786.

Saib, Mahdi-Salim, Julien Caudeville, Florence Carre, Olivier Ganry, Alain Trugeon, and Andre Cicolella. 2014. Spatial relationship quantification between environmental, socioeconomic and health data at different geographic levels. *International Journal of Environmental Research and Public Health* 11 (4): 3765–86. doi: 10.3390/ijerph110403765.

Schweers, Stefan, Katharina, Kinder-Kurlanda, Stefan Müller, and Pascal Siegers. 2016. Conceptualizing a spatial data infrastructure for the social sciences: An example from Germany. *Journal of Map & Geography Libraries* 12 (1):100–26. doi: 10.1080/15420353.2015.1100152.

Skinner, Chris. 2012. Statistical disclosure risk: Separating potential and harm: *Statistical Disclosure Risk. International Statistical Review* 80 (3):349–68. doi: 10.1111/j.1751-5823.2012.00194.x.

Sluiter, Roderick, Jochem Tolsma, and Peer Scheepers. 2015. At which geographic scale does ethnic diversity affect intra-neighborhood social capital? *Social Science Research* 54 (November):80–95. doi: 10.1016/j.ssresearch.2015.06.015.

Stansfeld, S. A., and M. Shipley. 2015. Noise sensitivity and future risk of illness and mortality. *Science of the Total Environment* 520 (July):114–9. doi: 10.1016/j.scitotenv.2015.03.053.

Statistical Offices of the Federation and the Länder. 2018. “Immigrant Rates. German Census 2011.” https://github.com/stefmue/georefum/blob/master/data/census.attr.rda.

Stimson, R. 2014. A spatially integrated approach to social science research. In *Handbook of research methods and applications in spatially integrated social science*, edited by R. Stimson, 13–25. Cheltenham, United Kingdom: Edward Elgar.

Termorshuizen, Fabian, Arjan W. Braam, and Erik J. C. van Ameijden. 2015. Neighborhood ethnic density and suicide risk among different migrant groups in the four big cities in the Netherlands. *Social Psychiatry and Psychiatric Epidemiology* 50 (6):951–62. doi: 10.1007/s00127-014-0993-y.

Tolsma, J., and T. W. G. van der Meer. 2017. Losing wallets, retaining trust? The relationship between ethnic heterogeneity and trusting coethnic and non-coethnic neighbours and non-neighbours to return a lost wallet. *Social Indicators Research* 131 (2):631–58. doi: 10.1007/s11205-016-1264-y.

Van den Eynden, Veerle, and Louise Corti. 2017. Advancing research data publishing practices for the social sciences: From archive activity to empowering researchers. *International Journal on Digital Libraries* 18 (2):113–21. doi: 10.1007/s00799-016-0177-3.

Vardigan, Mary. 2013. The DDI Matures: 1997 to the Present. 6.

Vardigan, Mary, Pascal Heus, and Wendy Thomas. 2008. Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation* 3 (1):107–13. doi: 10.2218/ijdc.v3i1.45.

Weßling, Katarina Dorothee. 2016. The influence of socio-spatial contexts on transitions from school to vocational and academic training in Germany. Empirical Research in Vocational Education and Training 7 (12). https://doi.org/10.15496/publikation-15222

Zandbergen, Paul A. 2014. Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine* 2014:1–14. doi: 10.1155/2014/567049.

Zenk-Möltgen, Wolfgang. 2012. Metadaten und die Data Documentation Initiative (DDI). In *Langzeitarchivierung von Forschungsdaten: Standards und disziplinspezifische Lösungen*, eds. Reinhard Altenhöner and Claudia Oellers, 111–26. Berlin: Scivero Verlag.

Zwickl, Klara, Michael Ash, and James K. Boyce. 2014. Regional variation in environmental inequality: Industrial air toxics exposure in U.S. cities. *Ecological Economics* 107 (November):494–509. doi: 10.1016/j.ecolecon.2014.09.013.